# Conclusiones reunión con Toñi sobre related work

Hola chicas:

Os resumo más abajo conclusiones de mi reunión de esta mañana con TOÑI… que casi ha sido en una cueva pues con la tormenta apenas había luz natural 

TOÑI: hemos quedado en que te encargarás de buscar propuestas más recientes en el campo de (multi-) relational data mining y en el campo de graph data mining.  Por favor, intenta pillar las principales (mirando la revista o conferencia) de los últimos 5 años y después tiramos para atrás si fuera necesario.  En cuanto lo tengas, me avisas y nos reunimos para sacar conclusiones.

PATRICIA: como comentamos anoche, tú me envías el primer draft con la primera descripción de ROLLER en forma de enumerado de ideas, para repasarlo antes de atacar la escritura del texto.  Por favor, como tú vas a llevar la batuta en este artículo, mete ya una sección con estas conclusiones y las de mi mensaje anterior (incluida la repuesta que tú mandaste), de forma que no se nos olvide nada de lo que hemos estado discutiendo.

Seguimos palante y trabajando a tope.  Saludos. ¡Buen finde! RC

**-- RESUMEN DE LA REUNIÓN -----------------------------------------------------------------------------------**

LO QUE HEMOS HECHO:

* Hemos visto en un libro que usan como baseline el dataset original, sin interpretar las características relacionales.  Quizá podría ser interesante que nosotros pusiéramos también las medidas que se obtienen con el primer dataset, antes de empezar a interpretar las características relacionales.  Así podríamos ver cuánto se mejora desde el primer paso hasta el final.  De momento apunto tan sólo esta idea para tenerla en cuenta en el futuro.
* Parece que las técnicas relacionadas en el campo de Machine Learning se agrupan en tres campos:
  + XML Mining:
    - Las hemos descartado pues todas están orientadas a encontrar reglas de asociación entre los nodos de los árboles XML.
  + Multi-relational data mining
    - Al parecer es lo mismo que relational data mining.  El “Multi”  es un prefijo que los autores quitan o ponen en función a sus preferencias.
    - Los trabajos que hemos visto en este campo intentan adaptar técnicas ya conocidas en el campo de la inducción de árboles de decisión o reglas para que sean capaces de aprender a partir de datasets expresados como “relaciones”, es decir, como predicados lógicos de primer orden con símbolos de constante tan sólo.
    - Hemos analizado SCART, que es una variante de SCAR para aprender árboles de decisión:
      * SCAR es para aprendizaje proposicional.
        + Toma el conjunto de entrenamiento y calcula todos los selectores de la forma A \theta v, en donde A es un atributo, \theta es un comparador (=, <>, >, <, >=, <=) y v es un valor posible de ese atributo.
        + Calcula un score para cada uno de esos selectores.
        + Se queda con el selector que tiene mejor score.
        + Al aplicar ese selector sobre el training set original se generan dos conjuntos: uno con los datos que cumplen el selector y otro con los  que no lo cumplen
        + Se vuelve a aplicar el mismo algoritmo a cada uno de los conjuntos, hasta que en una de las subdivisiones se obtiene un conjunto en el que todos los datos tienen la misma clase.
      * SCART es para aprendizaje “relacional”… pero en el sentido de primer orden
        + Básicamente es lo mismo, lo que ocurre es que en lugar de trabajar con selectores trabaja con los predicados de soporte.
    - También hemos analizado TILDE que es una variante de C4.5 para aprendizaje relacional.
      * Es la misma idea de C4.5, pero en lugar de usar selectores, que es lo normal en aprendizaje proposicional, utiliza predicados de soporte
  + Graph data mining
    - No hemos visto aún estos trabajos.

LAS CONCLUSIONES:

* Lo que llaman relacional es realmente aprendizaje de primer orden.
* Las propuestas que hemos estudiado toman inspiración de un algoritmo clásico de aprendizaje proposicional y extenderlo cambiando el language bias; es decir, en lugar de usar selectores de la forma A \theta v, utilizan predicados de soporte.
* A falta de estudiar las propuestas de graph mining parece que nuestros puntos fuertes son los siguientes:
  + El resto de propuestas son ad-hoc:
    - Toman como base una técnica existente, pero tienen que adaptarla para cambiar el language bias.
    - Nosotros tomamos cualquier técnica existente sin hacer ninguna modificación en la misma y la extrapolamos para que haga aprendizaje relacional.
  + El resto de técnicas parten de una representación en lógica de primer orden con símbolos de función tan sólo.  No hacen una distinción explícita entre las características atributivas y relacionales.  ´
    - Esto hace que el procedimiento de exploración tenga que trabajar con cada “relación” (predicado de soporte) de forma separada.  Es decir, es un procedimiento de exploración estilo FOIL.  En el caso de TILDE se usa look-ahead para mitigar un poco los efectos de la miopía, pero dicen que esto incrementa el coste computacional.
    - Nosotros analizamos todas las características atributivas de golpe y de forma tan eficiente como el algoritmo de aprendizaje base sea capaz de hacerlo.  Si el clasificador que obtenemos no es suficientemente bueno, entonces usamos una cantidad de contexto no acotada para mejorarlo (el contexto son los datos a los que se llega a través de las características relacionales).
* Si después de este análisis llegamos a la conclusión de que la técnica de ROLLER es original, tendríamos que pensar en venderla en un contexto más general:
  + La primera idea es en el contexto del Machine Learning
    - El problema aquí  es que para hacerlo tenemos que conseguir “vectorizar” los problemas de ILP clásicos.
    - De momento tengo algunas ideas, pero no logro amarrarlas suficientemente bien.
  + Se me ocurre una idea mejor en el campo de la Web Semántica.
    - Aquí los problemas se representan en Description Logic, que es una forma restringida de First-Order Logic en la que todos los predicados son unarios para indicar que algo es de algún tipo o binario, para representar propiedades de datos (attributive features) y de objeto (relational features)
    - Es decir, nos lo han puesto sencillito, sencillito dado que la representación es justamente la que nosotros necesitamos.
    - Y en el campo de la web semántica esto de aprender a partir de los datos es ciencia ficción, así que creo que podemos tener posibilidades de colarlo en una buena revista sobre el tema.  

--

Dr. Rafael Corchuelo

ETSI Informática

Avda. Reina Mercedes, s/n

Sevilla 41012

Tel.: +34 95 455 27 70

Fax: +34 95 455 71 39

Web: <http://www.tdg-seville.info/corchu>

Seeking for a paper?  Try TDG Scholar!

<http://scholar.tdg-seville.info>

Seeking for a listening platform? Try Opileak!

<http://www.opileak.com>

Seeking for integration?  Try Guaraná Solutions!

<http://www.guaranasolutions.com>